



情報学研究データリポジトリ IDR のご紹介

国立情報学研究所
データセット共同利用研究開発センター
大須賀智子，大山敬三

IDR：情報学研究データリポジトリ



情報学研究用データセットの共同利用事業

- 2010年より運用開始
- データセットの収集，受入，配布
：セミクローズドなデータの提供

⇒ コミュニティを繋いで研究を活性化

IDR発足の背景

データを使いたい
研究者

ビッグデータ, AI ... ⇒ 研究と実応用が接近

【従来】 人手で集めたデータ, 実験室で作ったデータ

【現在】 実サービスや社会で発生するリアルで大量のデータ

データを持つ
産業界

アカデミアへのデータ提供の動機はある

社会貢献, 現状の課題解決, 人材の確保, ...

- 大学と共同研究 → ツテが必要, 個別対応は大変
- 自社ビジネスへのリスク?

オープンデータにできないリアルなデータ (個人情報, 著作権, ...) を, 適正な管理の下で, 契約に基づき利用可能に

IDRで提供中の民間企業のデータセット

- Yahoo! データセット
- 楽天データセット
- ニコニコデータセット
- リクルートデータセット
- クックパッドデータセット
- LIFULL HOME'Sデータセット
- 不満調査データセット
- Sansanデータセット
- インテージデータセット
- オリコンデータセット
- ダイエットロコミデータセット
- 弁護士ドットコムデータセット

情報学研究データリポジトリ (IDR)

情報学研究データリポジトリ (Informatics Research Data Repository : IDR) は、国立情報学研究所 (NII) のデータセット共同利用研究開発センター (DSC) が運営するデータセットの共同利用事業です。IDRでは各種のデータセットを民間企業や大学等研究者から受け入れて研究者に提供するためのサービスを行っています。

データセットの利用を希望される方は、「[データ一覧](#)」から各データセットのページにアクセスし、記載内容をよくご確認ください。データセットの利用を申請された場合、「[IDRデータセット提供サービス規約 \(法人利用専用\)](#)」(データセットによっては「[同 \(個人利用専用\)](#)」) に同意いただいたものと見なします。

データセットの提供を希望される大学等の研究者の方は「[大学等研究者提供データセット受入](#)」のページをご覧ください。また、民間企業等の場合は、「[問い合わせ](#)」に記載のIDR事務局にご相談ください。

お知らせ

- 「[弁護士ドットコムデータセット](#)」の提供を開始しました。→ [プレスリリース 2020-03-11 NEW!!](#)
- 音声コーパス「[三重大学 日本語対話コーパス](#)」の提供を開始しました。→ [NIIトピックス 2020-03-10 NEW!!](#)
- 「[立命館ARC所蔵浮世絵データベース](#)」の提供を開始しました。→ [NIIトピックス 2020-03-05 NEW!!](#)
- 「[楽天データセット](#)」を更新しました。→ [NIIトピックス 2020-02-25](#)
- 「[NTCIRテストコレクション](#)」に「[NTCIR-14 FinNum](#)」を追加しました。2020-01-17

履歴

- 2019-11-29 「[IDRユーザフォーラム2019](#)」を開催しました。
- 2019-10-29 「[ダイエットロコミデータセット](#)」の提供を開始しました。→ [プレスリリース](#)
- 2019-09-26 「[大学等研究者提供データセット受入](#)」のページを公開しました。→ [NIIトピックス](#)
- 2019-09-02 「[インテージデータセット](#)」の新規データの提供を開始しました。→ [NIIトピックス](#)
- 2019-08-20 「[グループコミュニケーションコーパス \(TDU-NEDO\)](#)」の提供を開始しました。→ [NIIトピックス](#)
- 2019-07-05 「[楽天データセット](#)」に新規データを追加しました。→ [NIIトピックス](#)
- 2019-06-24 「[オリコンデータセット](#)」の提供ページを公開しました。→ [プレスリリース](#)

※2020/3 新規提供開始！

例：LIFULL HOME'Sデータセット

住宅情報サイト「LIFULL HOME'S」に2015年9月時点で掲載の
全国約533万件の賃貸物件データとサムネイル画像約8,300万枚



- 賃料，管理費
- 敷金，礼金，更新料
- 面積，間取り
- 立地（市区町村）
- 最寄り駅，徒歩分
- 築年数
- 建物構造
- 諸設備 など

※間取り図は高精細度の
画像データも追加提供

例：弁護士ドットコムデータセット ※3月提供開始！

オンライン法律相談サービス「みんなの法律相談」に投稿された約25万件の質問とそれに対する弁護士の回答のテキストデータ

- 質問のカテゴリ
- 「弁護士が同意」の人数
- 質問者が付与した「ありがとう」「ベストアンサー」

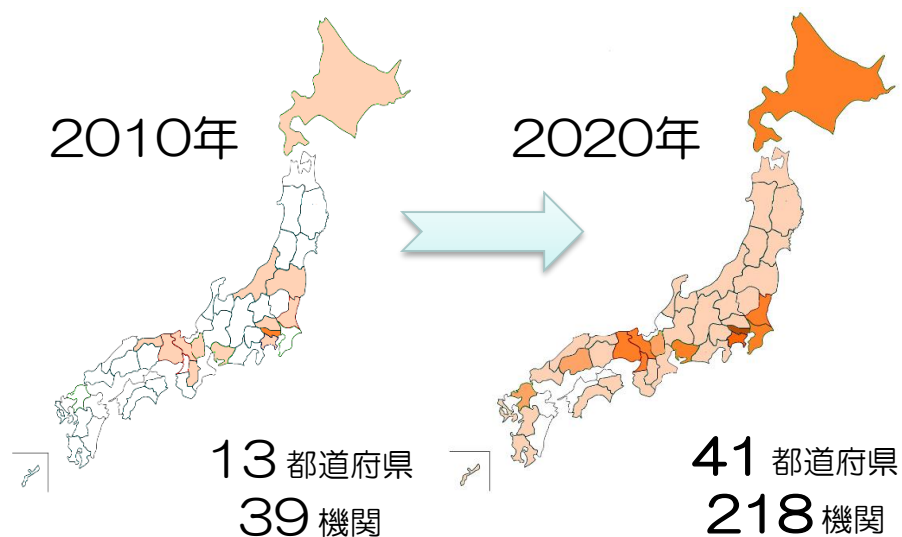


<https://www.bengo4.com/> のトップページより引用

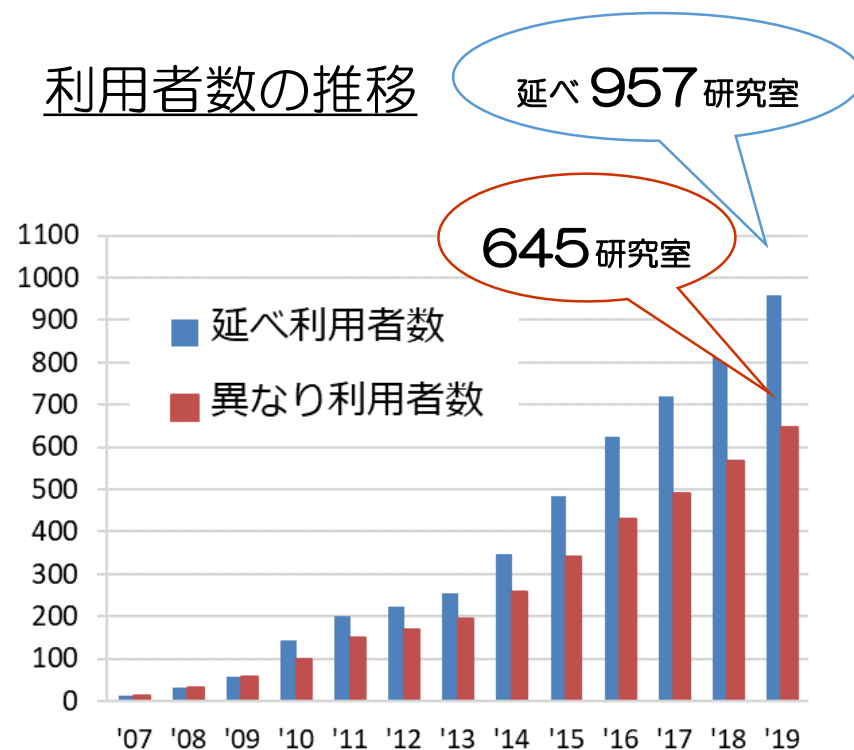
企業データ提供実績：研究室単位（～2020.3）

データ提供先

（ニコニコ，Sansanデータセット除く）

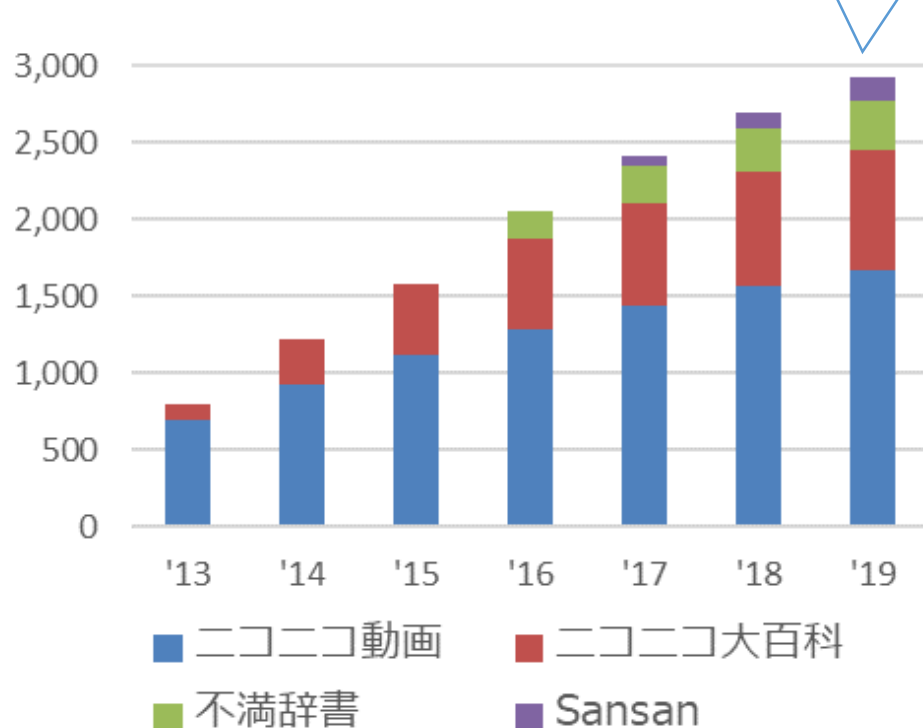


利用者数の推移

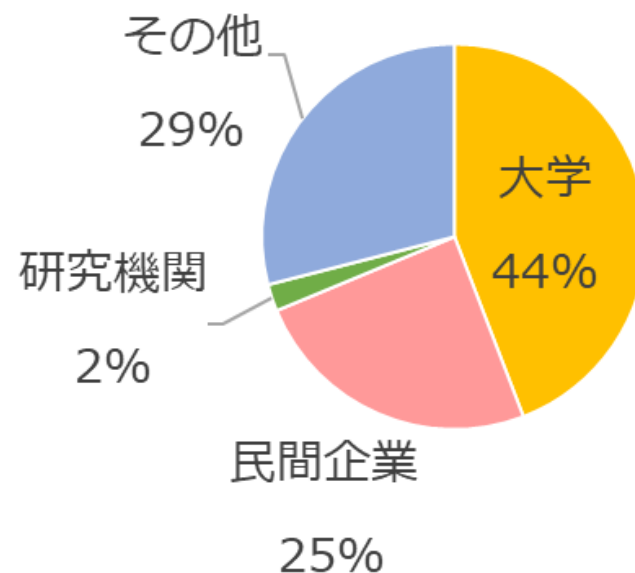


企業データ提供実績：個人単位（～2020.3）

利用申請者数の推移



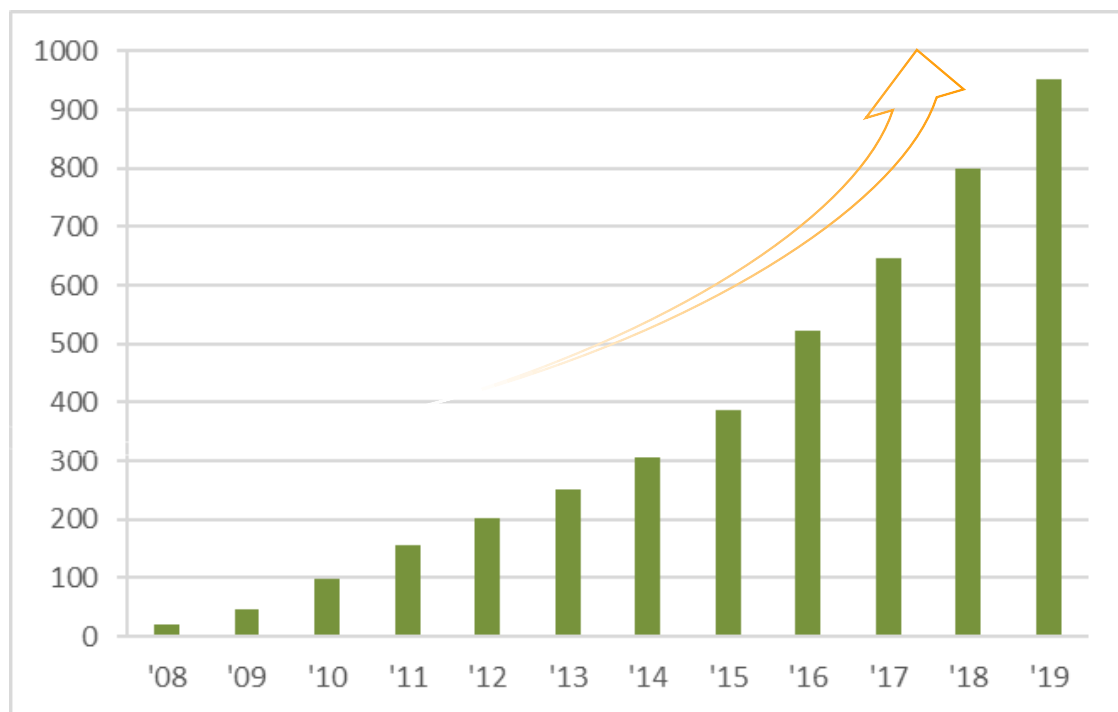
利用者の所属



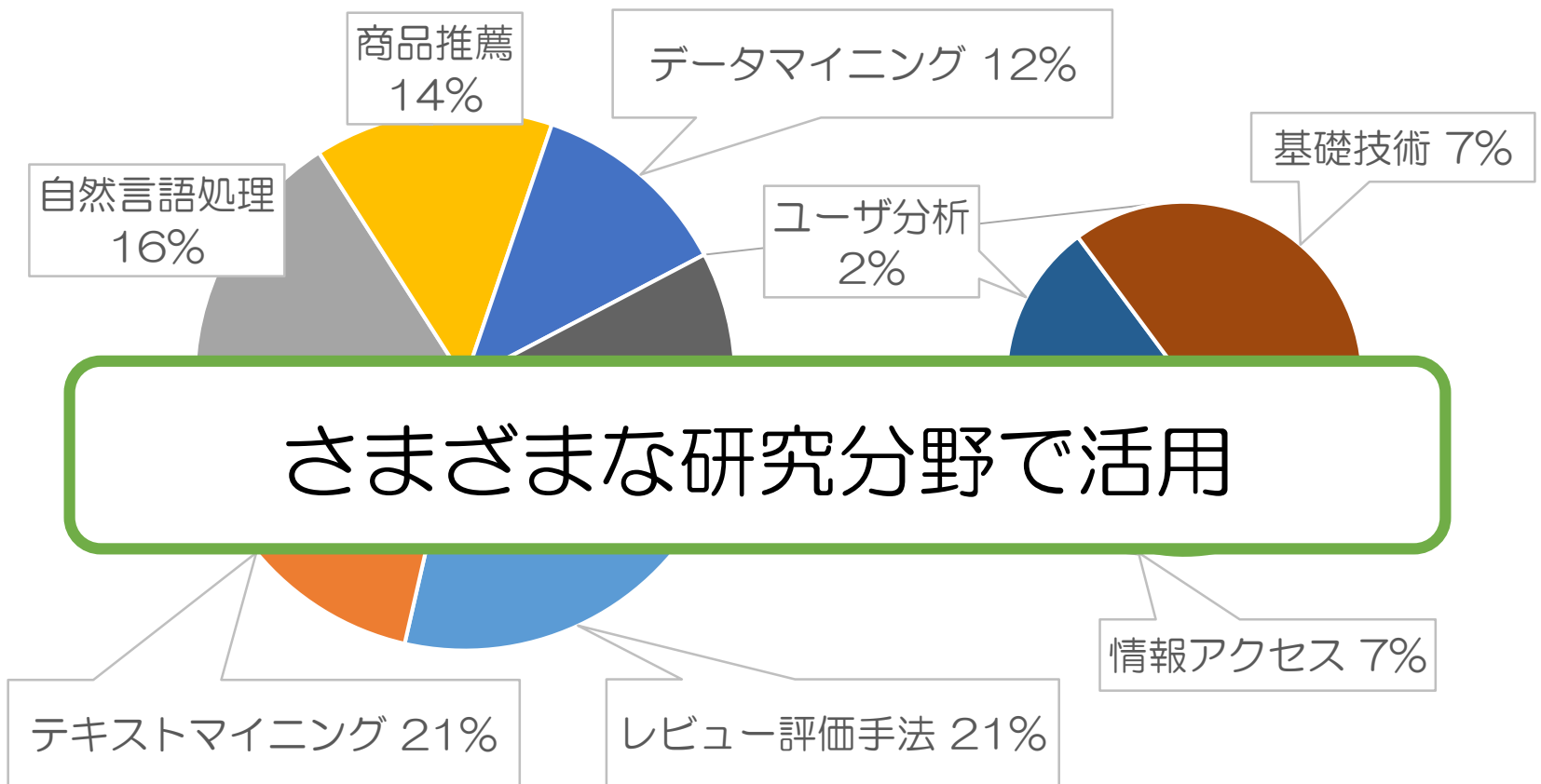
企業データを利用した研究発表（～2019年度）

研究成果発表数の推移

※毎年度の利用報告ベース



楽天市場データを利用した研究の例



データを利用した論文リストは公開しています

「研究発表」

インデックスリスト

- Yahoo!データセット [226件]
- 楽天データセット [219件]
- ニコニコデータセット [33件]
- リクルートデータセット [9件]**
- クックパッドデータセット [137件]
- LIFULL HOME'Sデータセット [25件]
- 不満調査データセット [26件]

アイテムリスト

研究発表 ▶ 情報学研究データリポジトリ (IDR) ▶ リクルートデータセット

チェックしたアイテムをExport 実行 表示順 出版年 (降順) 表示数 20

- ☐ ランダムフォレストを用いたカスタマーレビューの半自動分類
小暮 枝聖子, 倉藤 史哲, 石津 昌平
インテリジェントシステムシンポジウム2017予稿集, 154-157 (2017)
- ☐ 定量的分析を用いた自作自演の口コミ判定モデルの構築と評価
谷 芹菫, 藤原 智也, 波多野 賢治
IDRユーザフォーラム2017, (2017)
- ☐ レビュー投稿者の満足度を考慮した評判情報の可視化と店舗検索への応用
日高 加奈, 倉田 哲也, 大原 則三
IDRユーザフォーラム2017, (2017)
- ☐ 価値観モデルを適用した協調フィルタリングによるハイブリッド型推薦手法

研究者等提供データセットの受入を開始

- 受入要項を公開

- 流れ：

受入依頼書の提出



NIIで審査・手続き



提供同意書を提出

※本日提供開始！



NII 国立情報学研究所 National Institute of Informatics

情報学研究データリポジトリ

HOME データ一覧 研究成果一覧 ユーザフォーラム 組織 関連リンク お問い合わせ

> HOME > IDRサービス

大学等研究者提供データセット受入

2019/09/26 作成

情報学研究データリポジトリ (IDR) では、大学等の研究者（作成者）が研究用に作成した各種データセットを他の研究者（利用者）へ提供するための支援サービスを行っています。

本サービスでは、著作権や個人情報保護など様々な理由によりオープンデータとしての公開が難しいデータセットでも、作成者と協力して利用者の範囲、利用目的や利用方法などの条件を定め、IDRを通じて利用許諾契約を締結した利用者にデータセットを提供することによって、できるだけ作成者の負担を軽減しつつ、多くの研究者がデータセットを利用できるようにします。

受入条件

学術研究の対象として高い価値を有し、相当数の利用者が見込まれるデータセットが、本サービスの受入の対象となります。本サービスはアーカイブではありませんので、保存を目的とした受入はしません。

法令上または研究倫理上の観点から、IDRを通じた提供に必要な条件を満たさない場合や、IDRの能力を超える設備や取扱いが要求される場合は、受入をできません。

詳細については下記「書類一覧」の「研究者等提供データセット受入要項」をご覧ください。また、不明な点がございましたら、お気軽にお問い合わせください。

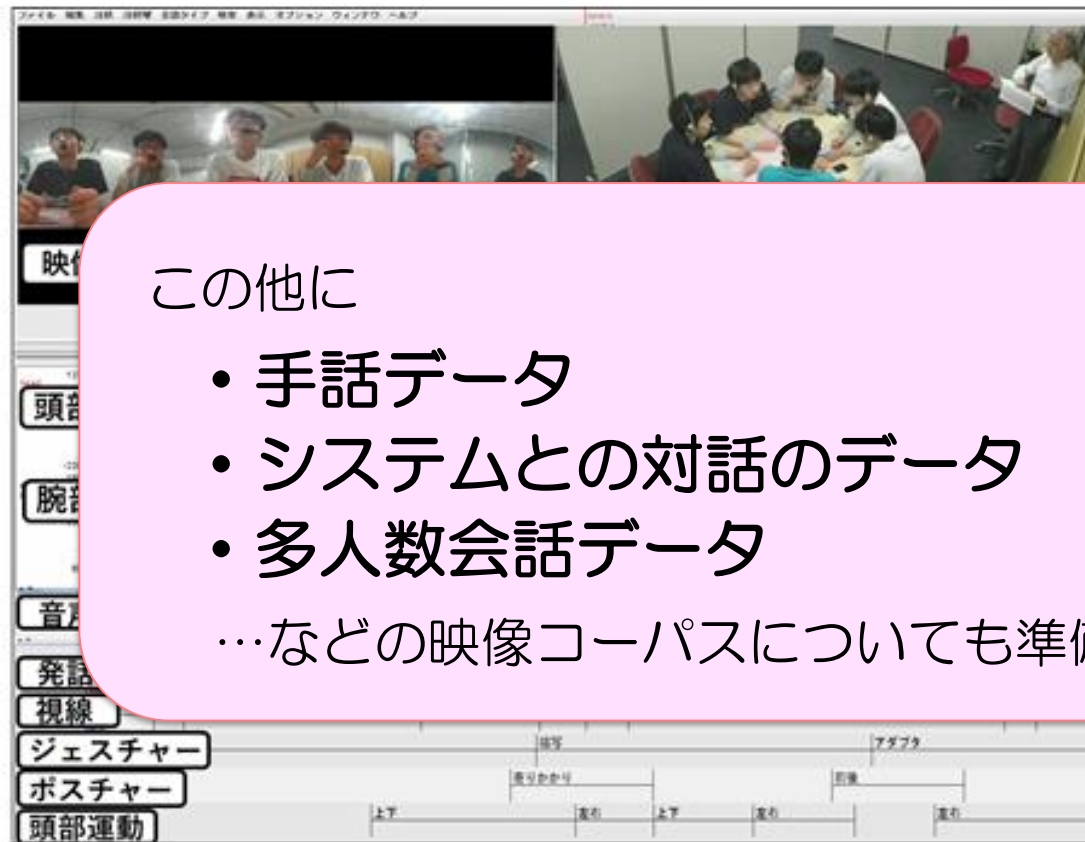
受入手続き

- 所属機関の契約担当者に、下記「書類一覧」の「研究者等提供データセット受入要項」第3条の「提供者」について確認してください。
※ 提供者は原則として研究代表者が所属する法人とします。ただし、例外として、当該法人によって適切であると判断された場合のみ、個人またはグループを提供者とすることができます。この場合、そのような判断の証となる資料の提出をお願いすることがあります。
※ データセット作成時とは所属法人が異なる場合や、作成したグループが複数の法人にまたがる場合などは、予めご相談ください。
- 「研究者等提供データセット受入要項」の内容に同意の上、下記「書類一覧」の「データセット受入依頼書」に必要事項を記入し、添付資料とともに、電子メールの添付ファイルで下記「問い合わせ窓口（IDR事務局）」に提出してください。
なお、様式は提供者に応じて「法人用」または「個人・グループ用」のいずれかを使用してください。
- 提出いただいた「データセット受入依頼書」に基づき審査を行います。審査の過程でデータセット受入依頼書の修正や追加資料の提出をお願いすることがありますので対応をお願いします。
- 審査に合格し、受入が決定した場合、「データセット提供同意書」（下記「書類一覧」のサンプルを参



研究者等提供データセットの例（1）

- グループコミュニケーションコーパス ※2019年8月提供開始！



研究者等提供データセットの例（2）

- 立命館ARC所蔵浮世絵データベース ※2020年3月提供開始！



浮世絵DB詳細情報表示 242件の中15件を表示中

クリックして画像のウィンドウを別々に開く → 縦A 縦B 縦C 縦D 縦E 縦F 縦G 縦H

同時作品表示 ← 前の作品 新規検索 後の作品 →

サムネイル一覧 簡易情報付一覧 ArtWiki English»

arcNo. arcUP3074 CoGNo. arcUP3069 浮世絵portal

絵師: 貞信 〈1〉 (Hasegawa Sadanobu:1) 落款印章: 一、雪花写 絵師検索

判型: 小判/錦絵 続方向: 横 作品位置: 027 枚組の 006 番目 01 枚続の 01 枚目

異版: Col重複: 1 All重複: 1 (撮影G: 頁: 1)

版元文字:

版印 1: - 版元 1: 版元名 1: -

改印: - 出版年月: ()・- 出版地: 太阪 同時作品

作品名: 「東海道五十三次 十一」 「東海道五十三次 十二」 「石薬師 四日市え二り半 # 丁」 「四日市 桑名え三り八丁」

配役: - 役者DB

系統分類: 名所絵 画題:

Ukiyo-e.ORG

所蔵: 立命館ARC 所蔵先管理No.: 資料部門: 浮世絵

Permalink: <http://www.dh-jac.net/db/nishkie/arcUP3074/default/>

主物・組み物 1280× 1024× 800× 100×

所蔵資料の利用に際しては、必ず利用ガイドをご覧ください。
If you wish to use image files not for personal, please click here.

Copyright © 立命館ARC. All Rights Reserved.

浮世絵の書誌情報
約11,000件
(画像データURL付き)

立命館大学アート・リサーチセンターの
サイトより引用

音声コーパス

音声資源コンソーシアム
Speech Resources Consortium

NII SRC
Speech Resources Consortium

ホーム SRCについて コーパスリスト 入手方法 資料室 FAQ リンク

ホーム > コーパスリスト

音声コーパスリスト

→音声コーパス検索システム
→音声コーパスを用いた研究成果一覧 (※DSC Reference Portal のサイトに飛びます)

無償コーパス

※当面は配布手数料(郵送料等)も無料とします

1. 重点領域研究「音声言語」・試験研究「音声DB」連続音声データベース (PASL-DSR)
2. 筑波大 多言語音声コーパス (UT-ML)
3. 東北大-松下 単語音声データベース (TMW)
4. 基盤研究(A)「日本語方言の地域差」方言音声コーパス (GSR-JD)
5. RWCP 音声データベース
 - a. 音声対話データベース - 96年版 (RWCP-SP96)
 - b. 音声対話データベース - 97年版 (RWCP-SP97)
 - c. 検索・要約用ニュース音声データベース (RWCP-SP99)
 - d. 会議音声データベース (RWCP-SP01)
6. RWCP 実環境音声・音響データベース (RWCP-SSD)
7. 重点領域研究「音声対話」対話音声コーパス (PASD)
8. CIAIR 子供の声データベース (CIAIR-VCV)
9. IPSJ SIG-SLP 雑音下音声認識評価環境 (CENSREC)
 - a. 雑音重畳日本語連続数字 音声認識評価環境 (CENSREC-1/AURORA-2J)
 - b. 雑音下日本語連続数字 音声区間検出評価環境 (CENSREC-1-C)
 - c. マルチモーダル音声認識評価環境 (CENSREC-1-AV)
 - d. 実環境車内 日本語連続数字 音声認識評価環境 (CENSREC-2)
 - e. 実環境車内 日本語単語 音声認識評価環境 (CENSREC-3)
 - f. 残響下日本語連続数字 音声認識評価環境 (CENSREC-4)

44種類のコーパス

- 読み上げ／講演／対話／3人会話
- 単語／短文／長文
- 成人／乳幼児／高齢者，非母語話者
- 方言，多言語，感情音声
- 雑音下，残響下， などなど…

15. 島崎詞音データベース91-92 (Tsuruoka)
16. 音声研究用X線フィルムデータベース (X-Ray)
17. 特定領域研究「韻律と音声処理」日本語MULTEX (MULTEX-J)
18. 中国語MULTEXコーパス (MULTEX-C)
19. 慶應義塾大学 研究用感情音声データベース (SENT-1)
20. 身体情報付き男・女・子どもの母音音素データベース (BIOVOW)
21. 東工大 多言語音声コーパス (TITML)
 - a. インドネシア語 (TITML-IDN)
 - b. アイスランド語 (TITML-ISL)
22. AWA長期間収録音声コーパス (AWA-LTR)
23. 南琉球新城方言音声データベース (Aragusuku)
24. 宮古大神島方言音声データベース (Oogami)
25. 感情評定値付きオンラインゲーム音声チャットコーパス (OGVC)
26. 千葉大学 3人会話コーパス (Chiba3Party)
27. 近畿大 児童の単語音声データベース (JWC)

有償コーパス

1. 日本音響学会 新聞記事読み上げ音声コーパス (JNAS)
2. 新聞記事読み上げ高齢者音声コーパス (S-JNAS)
3. 日本音響学会 研究用連続音声データベース (ASJ-JIPDEC)
4. NTT・東北大 親密度別単語了解度試験用音声データセット (FW03)
5. NTT・東北大 親密度別単語了解度試験用音声データセット2007 (FW07)
6. NTT 乳幼児音声データベース (INFANT)

IDRよりオンライン提供中

NTCIRテストコレクション



情報アクセス研究のためのテストベッドとコミュニティ

- NTCIRワークショップ
 - 1年半サイクルでタスクを実施
 - 共有テストコレクションを構築



- NTCIRカンファレンス
 - 各参加チームの成果を比較評価



- テストコレクションの公開

IDRより提供中

NTCIR-1	検索 ・学術 ・新聞 ・WEB ・地理情報 ・時間情報 ・音声文書 ・医療文書 ・数式 ・レシピ	用語抽出				
NTCIR-2			要約			
NTCIR-3				マイニング		
NTCIR-4		質問応答		多次元分類		特許検索
NTCIR-5						
NTCIR-6			動向可視化			
NTCIR-7				意見分析		特許翻訳
NTCIR-8		複雑な質問応答				
NTCIR-9			リンク発見			
NTCIR-10		入試問題				機械翻訳
NTCIR-11		含意・矛盾判断		意図理解		
NTCIR-12			モバイル検索			ライフログ

- タスクデータ33種類
 - 検索課題と正解判定
 - 質問と解答
- WEB文書データ2種類

(まとめ) IDR：情報学研究データリポジトリ



セミクローズドなデータの提供

- 民間企業データセット
- 研究者等提供データセット
- 音声コーパス
- NTCIRテストコレクション

実験環境の提供は無し

テストベッドとの
連携の可能性…？